LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods

Year: 2024 | Citations: 241 | Authors: Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou

---

Abstract

The rapid advancement of Large Language Models (LLMs) has driven their expanding application across various fields. One of the most promising applications is their role as evaluators based on natural language responses, referred to as "LLMs-as-judges". This framework has attracted growing attention from both academia and industry due to their excellent effectiveness, ability to generalize across tasks, and interpretability in the form of natural language. This paper presents a comprehensive survey of the LLMs-as-judges paradigm from five key perspectives: Functionality, Methodology, Applications, Meta-evaluation, and Limitations. We begin by providing a systematic definition of LLMs-as-Judges and introduce their functionality (Why use LLM judges?). Then we address methodology to construct an evaluation system with LLMs (How to use LLM judges?). Additionally, we investigate the potential domains for their application (Where to use LLM judges?) and discuss methods for evaluating them in various contexts (How to evaluate LLM judges?). Finally, we provide a detailed analysis of the limitations of LLM judges and discuss potential future directions. Through a structured and comprehensive analysis, we aim aims to provide insights on the development and application of LLMs-as-judges in both research and practice. We will continue to maintain the relevant resource list at https://github.com/CSHaitao/Awesome-LLMs-as-Judges.